DOCUMENT RESUME

ED 266 631 FL 015 445

AUTHOR Dickson, Peter
TITLE Evaluation of the Productive Skills in the Second Language: The Work of the Assessment of Performance Unit (APU).
PUB DATE 85
NOTE 15p.; In: Practice and Problems in Language Testing 8. Papers presented at the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (IUS) (8th, Tampere, Finland, November 17-18, 1984); see FL 015 442.
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adolescents; Evaluation Methods; Foreign Countries; Language Skills; *Language Tests; *Questioning Techniques; Secondary Education; *Second Language Learning; *Speech Skills; *Test Construction; *Test Use; *Writing Skills
IDENTIFIERS Great Britain

ABSTRACT

The work in second language education of the Assessment of Performance Unit, established by the Department of Education and Science of England and Wales to monitor certain areas of the public education curriculum is described. The focus is on the development of a rationale for testing, the elicitation procedures used in the assessment of speaking and writing, methods of assessing performance and their reliability, and the application of the results. Examples of tasks used in the productive language tests and some scales used to measure performance are appended. The methods were developed for testing second language students at the age of 13. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

EVALUATION OF THE PRODUCTIVE SKILLS IN THE SECOND LANGUAGE: THE WORK OF THE ASSESSMENT OF PERFORMANCE UNIT (APU)

Peter Dickson (National Foundation for Educational Research)

ED266631

1 Background

In carrying out the large scale testing necessary in a national survey -and the first APU survey of foreign languages involved about 12,000 pupils in schools in-England, Wales and Northern Ireland - many different factors have to be reconciled. The approach to the problems of test construction, assessment and interpretation needed to be pragmatic and not theoretical. The framework devised for the assessment does draw, of course, on language testing and language learning theory, but what I propose to describe in outline is the development of a rationale for testing, the elicitation procedures used in the assessment of speaking and writing, the methods of assessing performance, and the application of the results. There is not sufficient time here to give an account of the results being obtained. I shall restrict myself, therefore, to those aspects of the results which throw light on the effectiveness of the methods adopted.

2 Monitoring at age 13

I would like to say a word or two at this point about the background of the APU which was set up in 1974 by the Department of Education and Science in England and Wales. In the mid 1970s work began on the national monitoring of certain areas of the curriculum. Mathematics was monitored first, followed by English and then Science. The aim of the monitoring at the time was to obtain detailed information about pupil achievement in a situation in which individual schools made independent decisions about the content of the curriculum. The monitoring of foreign languages began in 1981. It was decided to monitor French, German and Spanish as a first foreign language in the first of the cycle of surveys in 1983 and to monitor French only in the subsequent surveys scheduled for 1984 and 1985. Another decision made at the time was to monitor pupils at age 13 only. It would not have been possible it was argued to obtain a representative sample nationally if the age chosen for monitoring had been greater, as many pupils are known to abandon studies of a foreign language at that stage. Equally, it was not thought profitable to monitor before that age as many pupils would have completed only two years' study by the age of 13.

At the beginning of the project the monitoring team had to consider the situation in schools. We also had to take account of the pattern of learning among pupils in the relevant age group. Teachers in schools had begun to seek new approaches to deal with the problems presented by the comparatively new situation in which pupils of all attainments were now learning a foreign language. This was one of the problems we had to confront in developing suitable instruments; the variety of teaching approaches used was another,and these problems were found to be especially acute when we came to assess language production. Opportunities for pupils to develop speaking skills in school were clearly limited. It was also claimed by many teachers that no writing in the foreign language

FL 015 445

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tomala

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

was done in the early stages, and some insisted that it was inappropriate to assess it.

3 Rationale for testing

Before the monitoring team was appointed, a working group set up by the APU had developed and published guidelines for the testing programme and the monitoring team took these as its first reference point. The predilection of the monitoring team for language tests with a communicative orientation was consistent with the proposals made by the working group. The guidelines set out a scheme thought relevant and this defined the linguistic content and sociolinguistic context of the kinds of test which they proposed. This meant that areas of the foreign language were specified in terms of topics, functions and general notions, and that the situations for the use of language were indicated in conjunction with the language skills required. No tasks were specified in the guidelines.

The priority, as we saw it, was to describe the testing domain by drawing up a range of settings and a list of tasks considered appropriate for the development of the tests. This had the obvious advantage of providing a rationale for identifying sources, both for texts to be used in the listening and reading tests, and for specific speaking and writing activities. This was a rationale which served also the interests of efficiency and economy, as it ensured that meanings focused on by the questions in the listening and reading tests were those which were implicit in the tasks, and that the speaking and writing activities were those which were most relevant in the chosen settings.

We have heard support in the course of this conference for testing discrete aspects of communicative performance. This explains our compromise position perfectly. To begin with we had no alternative but to accept the traditional four skill model of testing. Other approaches could not satisfy the rigorous requirements of measurement. Besides, the limited experience of language learning which characterised the population of 13 year old pupils in schools means that appropriate integrated tasks could not be identified in sufficient numbers for our purposes.

4 Framework for the writing and speaking tests

In order to carry out the monitoring, five tests in each of the areas of writing and speaking were developed. The five tests in each area are designed to cover a range of topics, and any one test is designed to cover a range of skills. Another feature of the construction which is common to both the writing and the speaking tests is the sequencing of tasks. At the beginning of each test, the tasks which pupils are required to carry out are closely structured. Subsequent tasks are progressively more open-ended. Some may argue that the final tasks still do not offer scope for a fully independent performance, but they should bear in mind the character of the target population.

BEST COPY AVAILABLE

In making decisions about the content and organisation of the tests, there were three principal considerations. First, we wished to satisfy ourselves that the tasks selected for inclusion in the tests were both useful and appropriate. That is to say that they were tasks which were accessible to the pupils both in terms of their personal experience as well as in terms of their language learning experience. Secondly, we needed to be sure that the tests were valid in terms of content. The skills required to carry out the tasks, and the topics, needed to be those which commonly formed part of the curriculum for 13 year old pupils. The third was a particular concern of the monitoring. The tests needed to be organised in such a way as to maximise the pupils' opportunities to demonstrate the range of skills which they possessed. Put another way, the tests needed to yield evidence of positive achievement at the lowest level, and yet impose no ceiling on levels of performance which might be found among 13 year olds. While performing this balancing act, we had also to make as much of each test as possible accessible to the lowest achieving pupils who should not, we thought, be alienated by the experience of taking part in the survey.

I should like to summarise the decisions taken to meet these concerns by describing the framework and content of the writing and speaking tests.

In the writing tests the tasks selected are centred on aspects of letter writing. Although other kinds of tasks such as message writing were considered, it was decided that letter writing was particularly suitable because it is an activity which is familiar to pupils in the age group, and also because it offers the flexibility necessary to introduce into the writing tests a range of topics.

Decisions as to which topics to include were less easy to arrive at, and were only settled after extensive rounds of pre-testing and a large scale pilot survey. In the final form of the tests there is a common component which is based on the topic of personal information, and another section specific to each test, which contains tasks based on one of a range of topics judged to be those with which 13 year old pupils are most familiar. Each writing test is in fact based on a single landscape picture which focuses on the topic, providing unity to the test as a whole.

How were we to maximise opportunities for pupils to write? How were we to avoid imposing an artificial ceiling on performance? One solution has already been referred to - the inclusion of both closely structured and open-ended tasks. Another solution was to ensure that pupils needed to draw on a reasonably wide range of skills in order to perform the different tasks included in the tests. Each writing test requires pupils to perform the skills of copying, listing, giving personal information, describing and reporting. The first of these, copying and listing, reflect activities which pupils commonly engage in in class and were included in the writing tests to ensure that the least able pupils have at least some tasks which are accessible to them. It is recognised that the decision to include these activities partly compromised the rationale which dictated that only realistic tasks or real life tasks should be included in the tests. It is relevant to note

BEST COPY AVAILABLE

here also that in order to elicit even the most modest attempts at writing, the tasks are presented in the test booklets in such a way as to encourage pupils to write something, even if what they write is incomplete or if they know they are making mistakes in the process of writing. It is not easy, of course, to know the extent to which that particular aim has been realised.

I turn now to the framework and content of the speaking tests. The tasks selected for inclusion in the speaking tests are drawn, as for other areas of the testing, from what were considered appropriate settings - travelling in a foreign country as a family, or in a school group, entertaining or visiting a person of a similar age from a foreign country. Within these settings, the tasks involve taking part in conversations, both social and transactional, and giving a more extended description or account. In order to accommodate this range of speaking tasks, each test includes a section requiring the pupils to read aloud, a closely structured conversation, two role tasks and a task requiring more extended speaking - a description or report of events. The point has already been made in respect of both the writing and the speaking tests that the organisation of the tests reflects progression from the closely structured to the open-ended. This is especially marked in the speaking tests in which the degree of interaction possible increases through the test, as also does the degree of independence required of the pupil.

As in the writing tests the content was settled as a result of extensive pre-testing and piloting of the material. In their final form the speaking tests have a common component, again on the pattern of the writing tests. This is the structured conversation which, in the interests of realism, or to mitigate its artificiality, has its own natural development, and a role scripted in French which needs to be played by the assessor. More relevant here is the content of the structured conversation which is designed to elicit many of the utterances which it is thought pupils might need to produce in initial encounters with foreign speakers. The linguistic content of the utterances cued by the role tasks is also to some degree prescribed. However, these are tasks in which a degree of interaction is possible. The inclusion of two role tasks in each of the speaking tests also ensures that there is a balance between the requirement to respond and to ask questions. The final task in each speaking test requires pupils to give a more extended account. Here the opportunity for pupils to demonstrate communicative strategies is supposedly at its greatest. Results, however, have shown lack of independence on the part of most pupils performing these tasks, compensated only by the answers to questions which the assessors are briefed to put to assist them.

I have given these details of the content of the writing and speaking tests, as illustration of the methods by itself would not convey an impression of the overall framework. I hope it will also provide a perspective for the description I shall give of the assessment criteria. It should be clear by now that one of our initial concerns in developing the tests was to satisfy the concerns of face validity. In devising the methods of assessment, we needed to take account not only of the rationale of the test construction, which placed tasks at the centre of the development, but also of the need to provide as much detailed description as was possible of pupils' performance.

BEST COPY AVAILABLE

5 Assessment criteria

The procedures devised for assessing performance of writing and speaking had to fulfil more than one function. They needed, on the one hand to provide overall scores, and on the other evidence of positive achievement at different levels. The solution was to develop scales of performance which provided not only grades for different levels of performance but also descriptions at each of the levels.

Below is an example of one scale used to assess the production of a written report of events.

5 Performs the task fully. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Performs the task adequately. Generally uses a range of appropriate structures and vocabulary. Inaccuracies do not obscure meaning.

3 Performs the task at a minimum level. Uses a limited range of structures and vocabulary, sometimes inappropriately. Inaccuracies may obscure some meanings.

2 Does not perform the task adequately. Uses a very limited range of structures and vocabulary, often inappropriately. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Nothing relevant to the task set.

- No attempt.

The descriptions were produced by reference to the broad criteria of communicative effectiveness, appropriateness and range. Effectiveness was considered to be the extent to which a pupil completed the task. Appropriateness included grammatical accuracy and coherence - phonetic accuracy and fluency in the case of speaking. Range related to the range of vocabulary and structure appropriate to the task. It soon became obvious in the development that a single scale of performance would not be applicable to the whole range of tasks. The differing nature of the tasks meant that the emphasis placed on any one criterion would not always be the same. For example, the parameters of the closely structured tasks were clearly set out by the separate cues, and range (or quantity) was inevitably a more crucial factor here in an effective performance.

The six points on the scale were arrived at by referring to the evidance of the pilot survey. This confirmed our view that six was a natural number of steps to include on the scale, and consideration of norms at that stage helped us to differentiate one from the other. Although the scales of performance provided descriptive categories, their main function was to facilitate rapid and reliable scoring. It is for this reason that the categories are economically expressed. It was found in the early stages of development that attempts to characterise

BEST COPY AVAILABLE

performance in detail by reference to a greater range of criteria tended to exclude large numbers of performances from classification. The simple description seemed in practice to be much more easily applicable. To summarise - the scales of performance are not arrived at by external standards but by reference to the variety of writing and speaking produced by the population surveyed. They are task-related. They provide broad descriptions of performance at each level. Most important, they provide a clear focus for distinguishing levels of performance. I shall return to the application of the scales for the purpose of scoring writing and speaking performance, but first I should like to complete my account of the procedures devised for carrying out the evaluation.

6 Detailed Analysis

In an attempt to elaborate on the grade descriptions, a system of detailed analysis of pupils' writing and speaking was worked out. Briefly - in the first survey the emphasis of the detailed analysis was on the components of the task performed, on the forms that were used to perform them and on the kinds of errors that were made. This kind of information was obtained on a selection of tasks, that is to say on one each of each type of task. The information was gathered in the following way. A list of items was entered on a grid by reference to a sample of the pupils' writing and speaking and the incidence of the use of each of these was entered by hand. The information obtained in this way provided the basis for producing detailed descriptions of performance of specific tasks. It was less easy however to use the information to make statements about different levels of performance generally. The approach to the detailed analysis will be further developed when the results of the 1984 survey are processed. Schemes which we are applying at present focus on other aspects of extended text creation in both writing and speaking. In writing we are aiming to obtain precise information about length, range and complexity, for example, in relation to levels of performance, and in speaking the schemes are directed in some cases towards the communicative strategies which pupils use in extended speaking, and in other cases, at the degrees of effectiveness and appropriateness pupils demonstrate in response to specific cues. This may help us to examine, in a systematic way, the difficulty pupils may have with different types of speaking and writing tasks. The national monitoring is of course a diverse exercise and, although we are collecting a rich set of data, we are conscious of the fact that it is not possible at present to do more than explore the possibilities of it in a general way.

7 Interpretation of Results

I would like to just summarise the uses to which the information obtained by the procedures I have described were put. First the unweighted scores obtained by placing performances on the scales I have described were aggregated to produce overall performance figures. These of course were crucial for reporting on performance in different modes of the assessment. Secondly, we were able to look at the distribution of scores across the scales of the performance both on individual tasks and on types of task for different groups of pupils. By reference to the evidence produced by the detailed analysis, we were able to elaborate descriptions of performance of specific tasks at different levels. Finally, we were able to review the evidence of the detailed analysis across the whole range of tasks in relation to the pattern of

BEST COPY AVAILABLE

distributions in order to produce generalised descriptions of performance at least at high, average and low levels. The starting point for these generalised statements were the scales of performance which were related to each other at the same level, even if they were articulated differently to take account of the different tasks.

8 Reliability

It is perhaps not appropriate to complete this account without referring to the measures we took to control the sources of variation present in any assessment of writing and speaking. It will be clear to you from what I have said that in developing the assessment materials a major concern was to demonstrate high content validity, and face validity. However, when we had completed the development of the tests with these concerns in mind, it became of crucial importance to investigate the question of reliability. In the course of pre-testing, and in the period between the pilot survey and the main survey, a number of measures were taken to increase the reliability of the tests. These improvements were brought about by reorganising the form of the tests and revising specific items, rewording instructions and so on. Results finally showed the rewards which are available if tests are painstakingly developed. In the pilot survey the reliability of some tests calculated by the Cronbach-alpha formula ranged from .3 to .9. In the main survey the reliability figures across all the tests, five in each of the areas of writing and speaking for French, German and Spanish, was in the narrower range of .85 to .9 and some were higher than that. We took this as evidence of a satisfactory degree of internal reliability. Indeed, it is hard to see how it could be improved without compromising the validity of the tests.

An equally pressing question of course is how to control for marker and inter-marker effects. The marking of speaking tests, in which the conduct of assessors is an additional variable, as we all know, presents the greatest problems. We are convinced that, following the proper development of the assessment materials, the most important task is the training of assessors and markers. The training carried out for the assessors of speaking needs also to be a feature of the training of the markers. They need to be completely familiar with the rationale and the content of the test material, and they need to understand the procedures laid down for the conduct of the tests. It is only at this stage that markers can benefit from practice in applying the scales of performance. They need to understand the concepts underlying the style of assessment represented by the scales of performance, and they need ample illustration of pupils' writing and speaking to exemplify performance at different levels. Practice in applying the scales should include discussions of differences. Our own experience is that this process is valuable in standardising the approach of assessors and in producing a high degree of consistency among teams of markers. Analyses of variance continue to bear out this view, even if random fluctuations persist in the marking of speaking.

It is of interest perhaps to mention one finding which emerged from these analyses. In the 1983 survey speaking assessors were required to score pupils' performance on-the-spot. It was found that the mean of these scores was lower than the mean score of markers working from tapes for every part of the speaking test. However, in spite of the fact that

BEST COPY AVAILABLE

assessors appear to mark with more severity their scores correlated highly with those of markers. Pearson correlations ranged from .78 to .81 for the four parts of the speaking test.

9 Conclusion

At the outset of this project we were conscious of the conflict which existed between the need to construct writing and speaking tests based on realistic tasks relevant to the current concerns in the curriculum, and of the need to satisfy the rigorous requirements of educational measurement. We were aware if you like of the need to effect a reconciliation between the concerns of validity and reliability in the language tests we were to use in the national monitoring. In this sense we hope at least to have shown that progress is possible by adopting a process of cautious compromise, realising what is possible of the innovative while safeguarding what is essential if judgements based on our assessment are to be taken seriously.

References

Department of Education and Science (1980)
Foreign Language. Assessment of Performance Unit.

Dickson, P., Boyce, C., Lee, B., Portal, M., Smith, M. (1985 in press)
Foreign Language Performance in Schools: Survey of French, German and Spanish. Report No. 1, London: HMSO.

Appendix 1. Examples of tasks used in APU speaking / writing tests, with some scales used to assess performance.

1. Writing tests
PART A. Giving personal information.

In a letter to your French penfriend you want to write about yourself and your family.
Use the list below to give as much information as possible.
Write complete sentences if you can.
If you can only write short phrases, do that instead.

Your penfriend will want to know:
What your name is.
How old you are.
If you live in the town or country.
If you have brothers or sisters.
What job your mother or father does.
What you like doing.
What you look like.
If you will write often.

Example: What you like doing. J'aime le football.
Write the information, in French, in the space below.

_____

_____

_____

....

BEST COPY AVAILABLE

PART B.

1. Listing

Look at the picture.
Below are two words for things in the picture.
Copy out the words into the spaces provided.

	Copy the words here
vaisselle	_____
parapluie	_____

Look at the picture again.
Now fill as many of the spaces below as you can, using words for things in the picture.
Write ONE word in each space.
Do not use words given above.

_____ _____

_____ _____

2. Describing

In a letter to your penfrind you want to describe the inside of your house.
Use the picture to help you decide what to write about.
Write your description using complete sentences if you can.
If you can only write short phrases, do that instead.
Write, in French, in the space below.

_____

_____

_____

....

3. Reporting

Last Saturday you decided to stay at home.
In a letter to your penfrind you want to say what you did during the day.
Write this, in French, in the space below.

_____

_____

_____

_____

...

BEST COPY AVAILABLE

PART B: Assessment scales

2. Describing

5 Describes several of the features/activities relevant to the task. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Describes some of the features/activities relevant to the task. Uses a range of appropriate structures and vocabulary. Many minor inaccuracies.

3 Describes some of the features/activities relevant to the task. Uses a limited range of structures and vocabulary. Inaccuracies may obscure precise meanings.

2 Refers to very few relevant features/activites using a very limited range of vocabulary. Structure recognisable. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Writes nothing relevant to the task.

- No attempt.

3. Reporting

5 Performs the task fully. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Performs the task adequately. Generally uses a range of appropriate structures and vocabulary. Inaccuracies do not obscure meaning.

3 Performs the task at a minimum level. Uses a limited range of structures and vocabulary, sometimes inappropriately. Inaccuracies may obscure some meanings.

2 Does not perform the task adequately. Uses a very limited range of structures and vocabulary, often inappropriately. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Nothing relevant to the task set.

- No attempt.

BEST COPY AVAILABLE

2. Speaking tests

PART A.

1. You are in a post office in France and you want to send this telegram to your penfriend. Read it aloud.

Destinataire	-	Vincent 9 rue de la Tannerie Poitiers
Toujours en vacances avec parents.		
Arrive lundi 13 heures gare SNCF.		

2. You are on a beach in France with your family. You want to find out from a French girl where you can buy an ice-cream. The assessor will play the part of the French girl.

1. Say hello.
2. Ask her if there is a café.
   She says there is, and she is going there too.
3. Ask her where the café is.
   She says it's at the end of the beach. You walk along together.
4. Ask her if she speks English.
   She prefers to speak French.

.....

12. Say thank you very much.
    She is going to leave the beach.
13. Say good-bye.
    She asks if you come to the beach every day.
14. Say yes, you will be there tomorrow at 10 o'clock.

PART A/2: Assessment scale

Structured conversation

5 Can convey the message in all cases, using expressions which a native speaker might use in the context.

4. Can convey the message in all or most cases, almost always using appropriate expressions.

3. Can convey the message in at least half of the cases, sometimes using appropriate expressions, at other times producing incomplete utterances.

2. Can convey the message in some cases, mostly in incomplete utterances. or:
   Can convey the message in few cases using appropriate expressions.

1. Can convey the message in a very few cases, and generally in incomplete utterances.

0. Nothing recognisable as relevant to the task.

- Not attempted.

BEST COPY AVAILABLE

PART B.
In this section of the test you take part in 2 short conversations. Read the instructions for each conversation and look carefully at the pictures. These will show you what you have to say.

1. Meeting a French boy

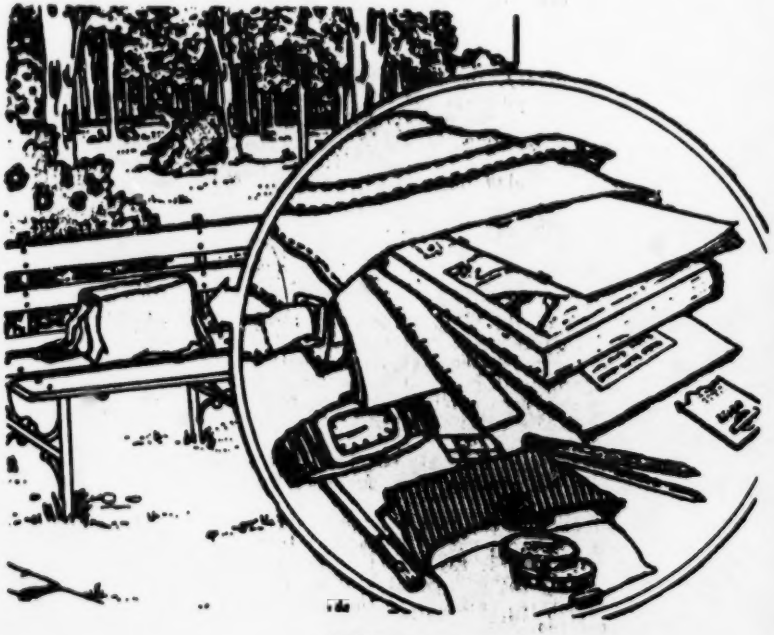
You meet a French boy at school who is on an exchange visit. You want to get to know him but he does not speak English. Find out his name and age and ask him about the different members of his family shown in the photograph below.

The assessor will play the part of the French boy.

2. Asking about lost property

You are on a school exchange in France and you find you have left your bag in the park. You go to the Lost Property Office to ask if it has been found.

The assessor will play the part of the person in the Lost Property Office.



BEST COPY AVAILABLE

PART B: Assessment scale

Rôle tasks

5 Conveys or elicits all details relevant to the task. In general, grammatically accurate.
4 Conveys or elicits all or most details relevant to the task. Uses some inappropriate forms. Some grammatical inaccuracy.
3 Conveys or elicits only details relevant to the task. Uses some inappropriate forms. In general grammatically inaccurate.
2 Produces language which succeeds in conveying or eliciting few details relevant to the task.
1 Says one or two words relevant to the task.
0 Says nothing relevant to the task.
- Not attempted.

PART C.

In this part of the test you have to give a more detailed description or account.

Make sure you have the sheet of colour photographs in front of you.

You want to describe to your penfriend a place where you have been on holiday. The photographs on the sheet are of the town, its surroundings and of the activities available. Use them to describe the place, what you can do there and where you can stay.

The assessor will play the part of your penfriend.

PART C: Assessment scale

Extended speaking

5 Uses a range of structures and vocabulary relevant to the performance of the task. In general grammatically accurate. Uses appropriate tense or tense indicators.
4 Performs the task adequately, using only one or two structures but an appropriate range of vocabulary. Makes some grammatical errors which do not obscure meaning. Some attempts to use tenses (or tense indicators) other than the present, if appropriate.
3 Performs the task at a minimum level. Structures and vocabulary are limited in range, and sometimes used inappropriately.
2 Does not perform the task appropriately. Produces some relevant language, but in fragmented form. Frequent grammatical errors. May resort to English.
1 Produces only one or two relevant words or phrases.
0 Produces nothing relevant to the task set.
- Not attempted.

BEST COPY AVAILABLE